Please cite this article as:
Haque, M. M., Chin, H. C. and Huang, H. "Modeling random effect and excess zeros in road traffic accident prediction." In proc. of 19th KKCNN Symposium on Civil Engineering, Kyoto, Japan, 2006.

Modeling Random Effect and Excess Zeros in Road Traffic Accident Prediction

*Md. Mazharul Haque[1], Chin Hoong Chor[2] and Huang Helai[3]

[123]*Department of Civil Engineering, National University of Singapore, Singapore-117576.*
mmh@, cvechc@, huanghelai@nus.edu.sg

[1] Research Scholar
[2] Associate Professor
[3] Research Scholar

ABSTRACT

Poisson distribution has often been used for count like accident data. Negative Binomial (NB) distribution has been adopted in the count data to take care of the over-dispersion problem. However, Poisson and NB distributions are incapable of taking into account some unobserved heterogeneities due to spatial and temporal effects of accident data. To overcome this problem, Random Effect models have been developed. Again another challenge with existing traffic accident prediction models is the distribution of excess zero accident observations in some accident data. Although Zero-Inflated Poisson (ZIP) model is capable of handling the dual-state system in accident data with excess zero observations, it does not accommodate the within-location correlation and between-location correlation heterogeneities which are the basic motivations for the need of the Random Effect models. This paper proposes an effective way of fitting ZIP model with location specific random effects and for model calibration and assessment the Bayesian analysis is recommended.

INTRODUCTION

In traffic accident prediction models, traffic accident rate or traffic accident cost is used as dependent variable. Other variables that are thought to provide information on the behavior of the dependent variable are incorporated into the model as predictor or explanatory variables. Needless to mention that regression analysis examines the relationship between a quantitative dependent variable and one or more quantitative or qualitative independent variables. The Poisson regression model is superior to multiple linear regression model to describe discrete, random, non-negative, sporadic accident data (Joshua et al. 1990). However, the Poisson distribution is an approximation of the binomial distribution for the situation that the number of trials is large and the success probability is low. This can be explained by the accident occurring process for a specific site. Dividing the year into 8760 one-hour periods, the chance that more than one accident will occur in any single hour is negligible and the occurrence of accidents is likely to be independent for the different hours. The hourly number of accidents would then be binomially distributed with *Binomial (8760, p)* since p is very low, where p is the probability of an accident in any given hour. This distribution is extremely close to the Poisson distribution with *Poisson (8760*p)*.

But there are several problems of Poisson regression. They are: 1) this model requires variance equals to the mean, but a number of studies (Miaou 1994, Shankar et al. 1995) found accident data to be over-dispersed, 2) if Poisson model is used in the presence of over-dispersion, estimated standard errors are likely to be very low (Cameron et al. 1986), 3) the accident rate is assumed constant for the sites with similar observed characteristics, but sites may be different in unobserved or omitted characteristics, thus leads to variance greater than mean, 4) in a Poisson process an event is equally likely at any point in a period of observation, but when an event occurs it increases the probability of further events in the same period, so this contagion property can be seen as a special cause of variation in the rate within a period of observation.

To overcome this over-dispersion problem, Negative Binomial (NB) model was found to be more suitable than Poisson model by introducing a stochastic component in the model (Miaou 1994, Shankar et al. 1995). This stochastic component relaxes the mean-variability constraint in the Poisson model. And this random error is assumed to be uncorrelated with independent variables and can be thought as the combined effects of unobserved variables that have been omitted from the model. So the NB regression model, which has the desirable distributional property to describe traffic accidents, is more general than the Poisson regression model. But location-specific effect is another factor that are not included in the NB model (Shankar et al. 1998), as a result serial correlation may occur due to multiple or repeated observations per location. Again another problem, faced by traffic accident prediction models, is the distribution of excess zero accident observations in some accident data. As a result simple Poisson or parent NB model estimates will be inherently biased because there will be an over-representation of zero-accident observations in the data, many of which do not follow the assumed distribution of accident frequencies (Shankar et al. 1997).

To handle the data structures with potential location-specific effects and excess zeros, the Random Effect model (e.g. MacNab 1998, Mitra et al. 2002) and Zero-Inflated count model (e.g. Chin et al. 2003, Miaou 1994) have been examined and increasingly applied for accident predictions in recent years. Most of these studies indicated the traditional models can be improved with the considerations of location-specific random effects and the dual-state system in zero-inflated model. However, a general model selection framework among those models is not available. And furthermore, it would also be interesting if a combination of ZIP and Random Effect model can further improve the performance of accident prediction model. The objective of this study is to explore an effective way of fitting Zero-Inflated Poisson (ZIP) with location specific random effects and propose the Bayesian analysis in the model calibration and assessment.

DEVELOPMENT OF PROPOSED MODEL

Random effect

Poisson and NB distributions are incapable of taking into account some unobserved heterogeneities due to spatial and temporal effects of accident data. In particular, in both of Poisson and NB model, it is presupposed that the accident occurrence distributions for the sites with similar observed characteristics are the same. Furthermore, accident counts for a specific location in different time periods are assumed to be independent with each other. But indeed, some hidden features may necessarily exist between different traffic sites and accident occurrences for a specific site may often be correlated serially. Consequently, without appropriately accounting for the location-specific effects and potential serially correlations, the estimates of the standard error in the regression coefficients may be underestimated. One way to overcome these problems is to treat them in a time series cross-sectional panel with different locations and time periods, as suggested by Hausman et al. (1984) in their study of patent applications. Using the panel data, the hidden features can alternatively be captured by individual (location) heterogeneity. The simplest random effects model for count data is the Random Effects Model (REP) that modifies the Poisson model as follows:

$$\lambda_{it} = L_{it}\alpha_i = e^{X_{it}\beta+\sigma_i} \tag{1}$$

where, λ_{it} is the modified Poisson parameter for random effects, L_{it} is the Poisson parameter representing the expected number of accidents on roadway location i in time t, α_i is the random location-specific effects assumed to be independently and identically distributed at the location level, $\mathbf{X}_{it}$ is the vector of covariates, $\boldsymbol{\beta}$ is a vector of estimate coefficients, and σ_i $(=\ln\alpha_i)$ is also the random location specific effects. According to Hausman et al. (1984), the modified Poisson probability specification as

$$\Pr(n_{it} \mid \mathbf{X}_{it}, \sigma_i) = \frac{\exp(-L_{it}e^{\sigma_i})(L_{it}e^{\sigma_i})^{n_{it}}}{n_{it}!} \tag{2}$$

where, n_{it} is number of accidents in roadway location i in time t.

Excess zero

Another challenge with existing traffic accident prediction models is the distribution of excess zero accident observations in some accident data. It is found (Shankar et al. 1997) that the distribution of annual accident frequencies may be qualitatively different from the simple Poisson and parent NB distribution because of the extra zero observations. To better reflect the situation, a dual-state system may be assumed. In this, one state is the zero-accident state, in which the traffic location, e.g. an intersection or a roadway section, can be regarded as virtually safe, while the other state is the non-zero-accident states, in which the accident frequencies are assumed to follow some known distributions such as the Poisson and NB. Since this zero-accident state may not seem valid with regard to all accidents, it can be an outgrowth of three sources: a) accident severity: minor accident may not be reported; b) near accident: it may also indicate a potentially dangerous traffic location even though no accidents have been recorded; (Shanker et al. 1997) c) specific types of accidents: some traffic location is possibly safe regarding to specific types of accidents (Chin et al. 2003). To handle this dual-state system, the Zero-Inflated Poisson (ZIP) model (Lambert 1992) has been developed.

$$\Pr(n_{it} \mid \mathbf{X}_{it}) = \begin{cases} p_{it} + (1-p_{it})\exp(-L_{it}) & n_{it} = 0 \\ (1-p_{it})\{\exp(-L_{it})L_{it}^{\,n_{it}}\}/n_{it}! & n_{it} > 0 \end{cases} \quad (3)$$

where, p_{it} is the probability of a roadway location i in time t to be in the zero-accident state or virtually safe. The overall probability of zero counts is a combination of the probabilities of zeros from each state which is represented in Eq. 3. Lambert (1992) has proposed that p_{it} be formulated as a logistic distribution such that

$$\log it(p_{it}) = \ln\left(\frac{p_{it}}{1-p_{it}}\right) = \boldsymbol{\theta}\mathbf{A}_{it} \quad (4)$$

where, $\mathbf{A}_{it}$ is the covariates vector for zero-accident state, $\boldsymbol{\theta}$ is estimated parameter vector.

The mean accident rate L_{it} in the non-zero-accident state satisfies a log-linear relationship with the covariates such that

$$\ln(L_{it}) = \boldsymbol{\beta}\mathbf{X}_{it} \quad (5)$$

where, $\mathbf{X}_{it}$ is the covariates vector for non-zero-accident state, $\boldsymbol{\beta}$ is estimated parameter vector for this state.

Random effect with excess zero

Although ZIP model is capable of handling the dual-state system in accident data with excess zero observations, it does not accommodate the within-location correlation as well as between-location heterogeneities which are the basic motivations for the need of Random Effect models. This paper proposes a combination of ZIP and Random Effect model to further improve the performance of accident prediction model. In particular, location-specific random effects can be considered into ZIP for both probability of being zero-accident state and count likelihood in non-zero-accident state. If these random effects truly exist, the Eq. 4 and Eq. 5 may lead to erroneous estimations of factor effects. Hence, these two equations may be rewritten as follows. For the zero-accident state, the equation is

$$\log it(p_{it}) = \ln\left(\frac{p_{it}}{1-p_{it}}\right) = \boldsymbol{\theta}\mathbf{A}_{it} + \psi_i \quad (6)$$

where, ψ_i is the random location-specific effect for zero-accident state. Due to some observed accident-inducing factors, it is reasonable to assume a correlation between different observations within specific site. Lambert (1992) also indicated that a slight variation in unobserved variables may cause the accident process to move back and forth between two states.

For the non-zero-accident state, the equation is

$$\ln(\lambda_{it}) = \ln(L_{it} * \alpha_i) = \boldsymbol{\beta}\mathbf{X}_{it} + \sigma_i \quad (7)$$

The modified Poisson parameter λ_{it} in Eq. 7 is also a random variable rather than a deterministic function of X_{it} as in Eq. 5. Correlation of λ_{it} and $\lambda_{it'}$ $(t \neq t')$ arising for different year t in a particular location i will be accounted for by α_i while λ_{it} and $\lambda_{i't}$ $(i \neq i')$ for different locations will be assumed to be independent as α_i is assumed independent.

MODEL CALIBRATION AND ASSESSMENT

In model calibration and assessment, either Maximum Likelihood Estimation (MLE) or Bayesian Inference (BI) may be used. In this paper, BI is employed using Markov Chain Monte Carlo (MCMC) algorithm. This choice of BI over MLE in accident analysis is important for several reasons when hierarchical data structure and extra zero observations are presented. Firstly, while in MLE, coefficients of factor effects are taken as fixed, BI appropriately represents the hierarchical data generating processes of accident occurrence by taking the parameters as unknowns with certain distributions (Gelman et al. 2003). Secondly, BI can accumulate evidence from any information sources regarding accident prediction. In Bayesian models, any engineering experiences or justified previous findings can be amounted into the posterior estimate of parameters by specifying the informative prior on those unknowns with preliminary information (MacNab 2003). Thirdly, since zero-inflated model could have multiple modes, the MLE are not always suitable for making inferences of parameters in this case while the Bayesian expected mean will be a better summary of the posterior than its modes (Angers et al. 2003). Moreover, as traffic accident record is sometimes incomplete, Bayesian approach provides a relatively easy way to bolt-on a missing data in the imputation procedure (Pardoe et al. 2006).

A model assessment using the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) both requires the specification of the number of parameters in the model, but in complex hierarchical models parameters may outnumber the observations and these methods cannot be directly applied. Therefore for model evaluation, Deviance Information Criterion (DIC), proposed by Spiegelhalter et al. (2003) is used. DIC provide a Bayesian measure of model complexity and fit that can be combined to compare models of arbitrary structure (Spiegelhalter et al., 2003). This can overcome the problems of classical criterions, such as AIC and BIC. Specifically, DIC is defined as:

$$DIC = D(\overline{\beta}) + 2P_D = \overline{D(\beta)} + P_D \tag{8}$$

where, $D(\overline{\beta})$ is the deviance evaluated at the posterior means of estimated unknowns $(\overline{\beta})$, and posterior mean deviance $\overline{D(\beta)}$ can be taken as a Bayesian measure of fit or "adequacy". P_D is motivated as a complexity measure for the effective number of parameters in a model, as the difference between $\overline{D(\beta)}$ and $D(\overline{\beta})$, i.e., mean deviance minus the deviance of the means. As a generalization of AIC, DIC can thus been considered as a Bayesian measure of fit or adequacy, penalized by an additional complexity term P_D. As with AIC, models with lower DIC values are preferred.

CONCLUSION

In this paper, Random Effect Zero-Inflated approach is proposed. As an explanation the Poisson model is developed to describe the accident data with location-specific effect and excess zero accident observations. This model can take into account the within-location correlations as well as between-location heterogeneities and also can support the dual-state system of accident occurrences for the distributions of excess zero observations in accident data. The Random Effect Zero-Inflated Negative Binomial model may better explain the accident data in case of greater individual-level variability. The development and derivation of this model will be presented elsewhere. For model calibration Bayesian inference is found to be more suitable than the Maximum likelihood estimation. And for complex hierarchical models, Deviance Information Criteria (DIC) is proposed for model assessment.

REFERENCES

Angers, J.F. and Biswas, A. (2003), "A Bayesian analysis of zero-inflated generalized Poisson model". *Computational Statistics and Data Analysis,* Vol. 42, 37-46.

Cameron, C. and Trividi, P.K. (1986), "Econometric models based on count data: comparisons and applications of some estimators and tests". *Journal of Applied Econometrics,* Vol. 1, 29-53.

Chin, H.C. and Quddus, M.A. (2003), "Modeling count data with excess zeros". *Sociological Methods and Research,* Vol. 32(1), 90-116.

Gelman, A., Carlin, J.B. and Stern, H.S. (2003), *Bayesian Data Analysis, 2nd edition*, Chapman and Hall, New York.

Hausman, J.C., Hall, B.H. and Griliches, Z. (1984), "Econometric models for count data with an application to the patents—R&D relationship". *Econometrica* Vol. 52 (4), 909–938.

Joshua, S.C. and Garber, N.J. (1990), "Estimating truck accidents rate and involvements using linear and Poisson regression models". *Transportation Planning and Technology,* Vol. 15(1), 41-58.

Lambert, D. (1992), "Zero-Inflated Poisson regression with an application to defects in manufacturing". *Technometrics,* Vol. 34, 1-14.

MacNab, Y.C. (2003), "A Bayesian hierarchical model for accident and injury surveillance". *Accident Analysis and Prevention,* Vol. 35, 91-102.

Miaou, S.-P. (1994), "The Relationship between Truck Accidents and Geometric Design of Road Section: Poisson Versus Negative Binomial Regression". *Accident Analysis and Prevention,* Vol. 26 (4), 471-82.

Mitra, S., Chin, H.C. and Quddus, M.A. (2002), "Study of intersection accidents by maneuver type". *Transportation Research Record*, Vol. 1784, 45-50.

Pardoe, I. and Weidner, R.R. (2004), "Sentencing convicted felons in the United States: a Bayesian analysis using multilevel covariates". *Journal of Statistical Planning and Inference*, Vol. 136, 1433-1455.

Shankar, V. N., Mannering, F. and Barfield, W. (1995), "Effect of Roadway Geometric and Environmental Factors on Rural Freeway Accident Frequencies". *Accident Analysis and Prevention,* Vol. 27 (3), 371-89.

Shankar, V. N., Milton, J. C. and Mannering, F. (1997), "Modeling Accident Frequencies as Zero-Altered Probability Process: An Empirical Enquiry." *Accident Analysis and Prevention,* Vol. 29 (6), 829-37.

Spiegelhalter, D. J., Thomas, A., Best, N. G. and Lunn, D. (2003). *WinBUGS version 1.4.1 User Manual*, MRC Biostatistics Unit, Cambridge, UK.